D394–D399 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 18 November 2011*
doi:10.1093/nar/gkr1019

ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins

Hong Luo[1,2], Ke Lin[1], Audrey David[1], Harm Nijveen[1,2,3] and Jack A. M. Leunissen[1,2,3,*]

[1]Laboratory of Bioinformatics, Wageningen University and Research Centre, P.O. Box 569, 6700 AN Wageningen, [2]Netherlands Bioinformatics Centre (NBIC), P.O. Box 9101, 6500 HB Nijmegen and [3]Netherlands Consortium for Systems Biology (NCSB), P.O. Box 94215, 1090 GE Amsterdam, The Netherlands

Received August 11, 2011; Revised October 19, 2011; Accepted October 21, 2011

ABSTRACT

ProRepeat (http://prorepeat.bioinformatics.nl/) is an integrated curated repository and analysis platform for in-depth research on the biological characteristics of amino acid tandem repeats. ProRepeat collects repeats from all proteins included in the UniProt knowledgebase, together with 85 completely sequenced eukaryotic proteomes contained within the RefSeq collection. It contains non-redundant perfect tandem repeats, approximate tandem repeats and simple, low-complexity sequences, covering the majority of the amino acid tandem repeat patterns found in proteins. The ProRepeat web interface allows querying the repeat database using repeat characteristics like repeat unit and length, number of repetitions of the repeat unit and position of the repeat in the protein. Users can also search for repeats by the characteristics of repeat containing proteins, such as entry ID, protein description, sequence length, gene name and taxon. ProRepeat offers powerful analysis tools for finding biological interesting properties of repeats, such as the strong position bias of leucine repeats in the N-terminus of eukaryotic protein sequences, the differences of repeat abundance among proteomes, the functional classification of repeat containing proteins and GC content constrains of repeats' corresponding codons.

INTRODUCTION

Amino acid tandem repeats, as one of the most prevalent patterns in protein sequences, have inspired the interests of researchers for many years in terms of their pathological, functional and evolutionary roles. According to the patterns of units, repeats in proteins can be generally classified into several categories.

Single amino acid repeats (SAARs), also known as homo peptides, have the simplest repeat unit. Some of the SAARs have been extensively studied as they are involved in numbers of human neurodegenerative diseases, such as those with variable polyglutamines (polyQ) and polyalanines (polyA) (1). Since they are important modulation factors on protein–protein interactions (2,3), the insertions, deletions, substitutions, as well as growing or shrinking of the repeats result in either loss-of-function or gain of abnormal function (4,5) by altering the conformation of protein tertiary structures. As for other types of SAARs, leucine runs are mainly located in the N-terminus of eukaryotic proteins, which are presumed to be involved in the signal peptide (6). Higher frequency of proline repeats in mammalian proteomes is considered to significantly contribute to network evolution (7). In addition, histidine repeats play a crucial role in the localization of human proteins to the nuclear speckle compartment (8).

Amino acid tandem repeats with complex unit patterns have also been studied frequently. Different from SAARs, most of them are comparatively conserved in their structure. Well-known patterns include the leucine rich repeats (LRRs) that commonly act as the structural framework for the formation of protein–protein interactions (9), the ankyrin repeats that contain the binding site for the huge titin proteins that are involved in muscle ultrastructure and elasticity (10,11), and the polyubiquitins that are synthesized as repetitive polyproteins (12).

Although the biological significance of particular amino acid tandem repeats have been demonstrated continually during the past years in several model organisms, no convincing conclusions can be drawn until now. The arguments are mainly posed on several aspects: Is the role of similar repeat patterns coherent in different proteomes across different life kingdoms? Could the functional and evolutionary roles of certain repeats correspond to their particular characteristics, such as position bias, GC content constrains and codon usage? How could the conserved functions of particular repeats have been evolved by natural selection? Why repeats are so

*To whom correspondence should be addressed. Tel: +31 317 482036; Fax: +31 317 418094; Email: jack.leunissen@wur.nl common in protein sequences even under the scenario that their instable characteristics are often relevant to disorder and diseases (5,13,14)? And what are the structural and sequence-based strategies (15,16) to prevent repeats from possible aggregation?

The dilemma of contradicting explanations of the role of repeats is partly because of the lack of repositories for large-scale investigation and comparison of repeats among the variety of proteomes across different kingdoms. Several databases of amino acid repeat were constructed during the recent decade. Unfortunately, some of these databases are no longer accessible or functional anymore such as COPASAAR (17), RepSeq (18) and ProtRepeatDB (19). As for the remaining ones, TRIPS gathered repeats generated from a very old version of SwissProt (year 1999) (20), RCPdb offers the codon usage bias data of homopeptides (SAARs) of 13 completely sequenced eukaryotic species (21), and the PolyQ database collects the sequences of all human proteins containing runs of seven or more glutamine residues (22).

To change the incompatible situation between the rapid increase of protein sequence data and the lack of a large scale, well-annotated protein repeat repository, we have constructed an online database of protein repeat sequences (ProRepeat, http://prorepeat.bioinformatics.nl/). ProRepeat recruits both perfect and approximate tandem repeats from all taxa of UniProtKB (23) and supplied by 85 complete sequenced and well annotated eukaryotic proteomes. ProRepeat also gathers the corresponding nucleotide sequences of the repeat fragments for the purpose of codon usage analysis. The latest update of ProRepeat is based on the datasets of UniProtKB release 2011_05 and RefSeq (24) release 40. An easy to use web interface was designed for users to query the database, and to perform statistical analyses on the query results. We believe that ProRepeat provides the user community with a useful resource for the exploration of function and evolution of protein repeats.

REPEAT DETECTION AND DATASET GENERATION

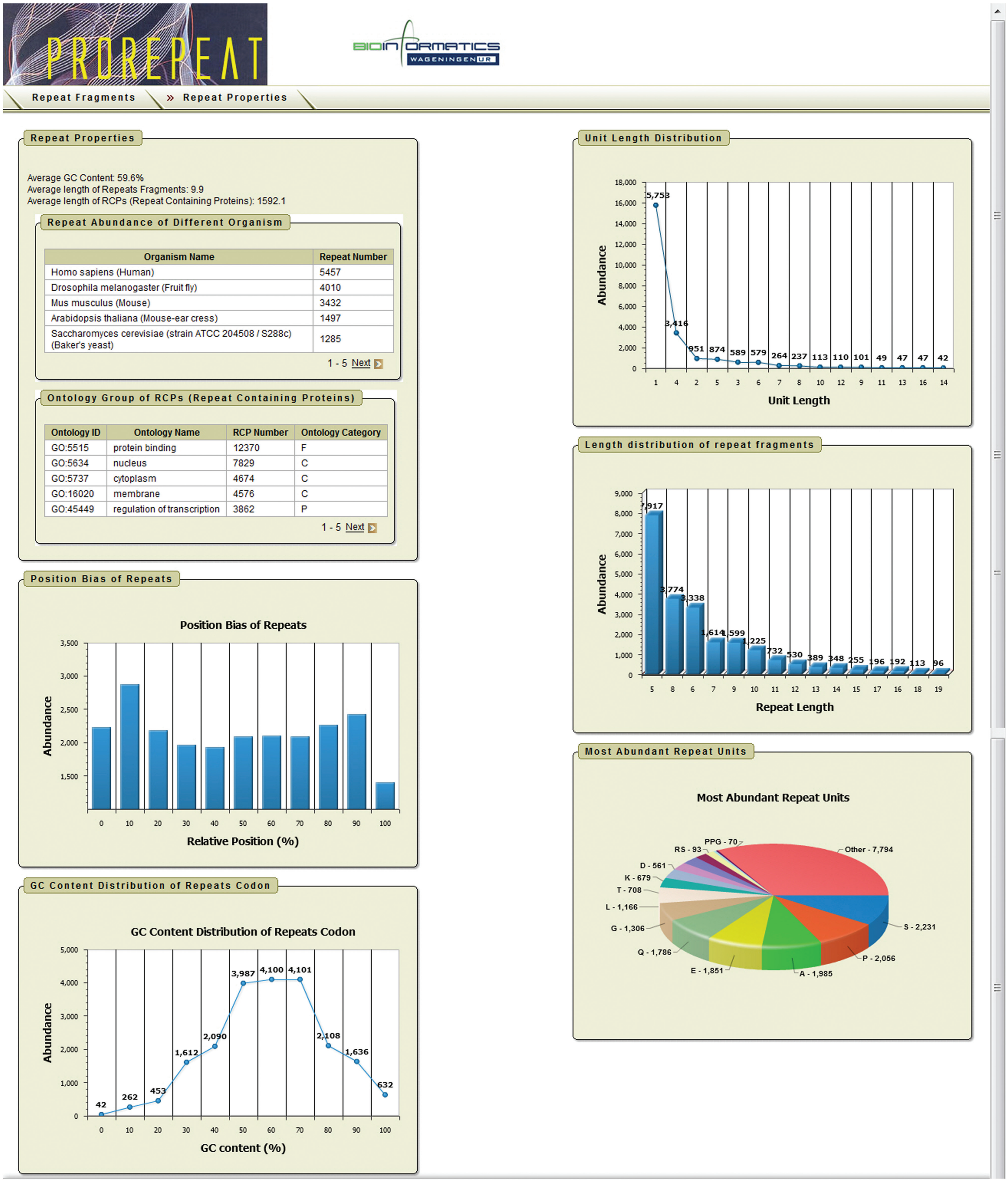
We collect three types of repeat patterns including perfect tandem repeats (PTRs), approximate tandem repeats (ATRs) and simple sequences (SSs) in proteins. The PTRs were detected using an in house developed C/C^{++} procedure we implemented based on the suffix tree algorithm which identifies all perfect tandem repeats in a protein sequence (25), the ATRs were detected by XSTREAM (26) and the SSs were detected by SIMPLE (27). Following the definition of statistically significant repeat runs in protein sequences (28), we used cutoff sizes of five, four, three and two of the repeat unit repetitions to identify mono-amino, di-amino, tri-amino and all other repeats, respectively.

It is possible that different algorithms identify the repeats with the same unit and overlapped position in the same protein. To integrate the repeats datasets excluding the redundancy, we developed a PL/SQL procedure which distinguishes between unique and overlapping repeats. The repeats datasets were merged together followed by a sorting step based on the identifier of repeat containing proteins (RCPs), repeat unit and the position of the repeat. The repeats with the same unit and overlapping position within the same protein were merged into a single fragment. If the begin and end positions of these repeats were also the same, only one of them was retained as they were actually the same repeat identified by different algorithms. We also classified the perfect and approximate repeats separately and marked them within the database, so that the user can search them individually.

The repeat datasets were generated based on the protein entries collected in UniProtKB release 2011_05. For the convenience of comparative analysis, we also generated the repeat datasets of the completely sequenced eukaryotic proteomes based on RefSeq release 40. For the selection of completely sequenced eukaryotic proteomes, we obtained the list of the complete published eukaryotic organisms from the genomes online database—GOLD (29). For each organism, we compared the number of ORFs given by GOLD with the number of proteins collected by RefSeq. If the two numbers were approximately consistent, i.e. the difference was <5%, we considered the proteome collected by RefSeq as complete and retrieved repeats from it. Thus, ProRepeat contains repeats from 85 complete sequenced eukaryotic proteomes including 14 vertebrates, 8 plants, 22 fungi, 12 insects and 29 other organisms. The gene ontology cross references of RCPs were generated based on GOA (30), and RefSeq annotations for gene ontology. The corresponding nucleotide sequences of the repeats fragments were obtained via EMBL and RefSeq cross-references within each UniProtKB and RefSeq protein entry, respectively.

THE WEB INTERFACE

The ProRepeat database can be accessed using an intuitive web interface. An Introduction page provides information about the types of repeats that the database contains, the tools that were used to create the database, and the background of functional studies of repeats. The Statistics page lists several characteristics of the database, like the abundance of different repeat types across the different life kingdoms. The Help page offers practical examples to help users find interesting repeats and perform online analyses. On the Query page, users can search for repeats in one or more species, using annotations of RCPs including entry ID, protein description and gene name. Users can also specify the repeat unit, unit length, repeat sequence length, number of units and position of repeat in the protein sequence. For the repeat unit, ProRepeat offers two additional options. For example, the repeat unit of two repeat fragments DEDEDEDE and EDEDEDED could be identified as DE and ED, and defined as isomorphic repeats. By switching on the 'Isomorphic Unit Search' option, users can obtain all cyclic permutations of this repeat pattern. With the 'ProSite Syntax Search' option, users can specify a regular expression as search pattern used by the ProSite database. In addition, ProRepeat classifies the repeats as



Average GC Content: 59.6%
Average length of Repeats Fragments: 9.9
Average length of RCPs (Repeat Containing Proteins): 1592.1

Organism Name	Repeat Number
Homo sapiens (Human)	5457
Drosophila melanogaster (Fruit fly)	4010
Mus musculus (Mouse)	3432
Arabidopsis thaliana (Mouse-ear cress)	1497
Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	1285

1 - 5 Next

Ontology ID	Ontology Name	RCP Number	Ontology Category
GO:5515	protein binding	12370	F
GO:5634	nucleus	7829	C
GO:5737	cytoplasm	4674	C
GO:16020	membrane	4576	C
GO:45449	regulation of transcription	3862	P

1 - 5 Next

Figure 1. The statistical analysis result of repeat properties for PTRs of all taxa in UniProt with the 'Isomorphic Search' option on, and the default evidence at protein level.

PTRs and ATRs defined by the similarities of repeat units. Users can choose either 'Perfect Tandem Repeats', 'Approximate Tandem Repeats' or 'All Repeats' (both PTRs and ATRs).

Users can query individual species of 85 eukaryotic complete proteomes obtained from RefSeq through their taxonomic names, or from broader taxonomic ranges collected from the UniProt Knowledgebase. The query results are displayed as interactive web pages in tabular format. Columns contain information about the position and length of the repeat, as well as the corresponding protein. Clicking on a repeat brings up a page with the corresponding DNA sequence and the codon usage pattern. Users can save the query results as Microsoft Excel format, or perform a secondary search using the query results. ProRepeat provides users with an online tool to perform statistical analysis of the query results. For example, the query results of PTRs from all species in UniProt at protein existence level can be analyzed to show various properties including the repeat abundance in different organisms, the gene ontology annotation of the RCPs, the position bias of repeats, the distributions of GC content of repeat codon, the unit length and repeat fragment length (Figure 1).

PRELIMINARY ANALYSIS OF PROTEIN REPEATS

Based on UniProtKB datasets, ProRepeat gathers ∼3.75 million repeat fragments contained in 2 million RCPs from 0.1 million organisms. The distribution of repeats over eukaryota, bacteria, archaea and viruses is shown in Table 1. The relative repeat abundance normalized by the number of proteins of the different kingdoms indicates that eukaryotic proteins are four times more likely to have tandem repeats than prokaryotic proteins, and the possibility of having tandem repeats in viruses and prokaryotes is similar. This supports the idea that large amounts of protein repeats arose after the divergence of prokaryota and eukaryota (31).

There is a long-standing debate about the roles of repeats in proteins. Some early viewpoints ascribe large amounts of SSs to 'junk protein' (32) as few of them have identified stable tertiary structures (33) and are thought to be non-functional. However, more and more evidences show that they are not just 'junk' peptide sequences (34) and might have particular function and structure (35). Important subsets of SSs, in particular cryptic and identical SAARs, have been reported to be actively evolving (36–38). As a result, the evolutionary footprint and functional implication of repeats which are being modulated by selection could be inferred from their properties. For example, in *Drosophila* and *Arabidopsis*, the RCPs are mostly involved in gene regulation, signaling and developmental processes, but significantly under-represented in the process of DNA recombination and DNA replication. In addition, the positional distribution of repeats in proteins of *Drosophila* and *Arabidopsis* is also non-random (39,40).

Using ProRepeat, we made a comparison of the repeat properties including repeat length, RCPs length, repeat position and repeat codon usage in model organisms across different kingdoms (Table 2). In general, glutamic acid (E), serine (S), glutamine (Q), proline (P), alanine (A) and leucine (L) are widely used by SAARs in all taxa, while the pattern varies between different taxa. For example, polyL and polyP are preferred by prokaryotes and eukaryotes, respectively; *D. melanogaster* uses polyQ more frequently than most of the other organisms; polyE is extremely abundant in Hepatitis delta virus, and for human immunodeficiency virus, although polyE has the highest frequency, when adding the approximate SAARs together arginine (R) is actually the most commonly used amino acid (near 80%). When looking at the N-terminal perfect SAARs, polyL is the most popular especially in eukaryotes and bacteria, in which they play functional roles, for instance, in signal peptides (6).

Table 1. Repeat abundance in four kingdoms

Kingdom	Repeat number		Repeat abundance[a] (%)	Protein abundance[b] (%)	Relative abundance[c]
	PTR	ATR			
Eukaryota	1 163 368	1 195 655	63.10	27.2	2.32
Bacteria	498 071	705 575	32.20	63.9	0.50
Archaea	12 584	18 631	0.85	1.8	0.47
Viruses	75 109	68 821	3.85	6.9	0.56

[a] Percentage of repeat numbers in four kingdoms, [b] Percentage of protein numbers in four kingdoms based on UniProtKB (0.2% unclassified entries are not listed), [c] Percentage of protein abundance divided by percentage of repeat abundance.

Table 2. Repeat properties in representative species

Species	Most abundant SAARs(%)	N/C SAARs	GC1	GC2	L1	L2
HIV	E(45.3), A(27.0), N(8.6)	SA/INP	42.0	41.9	462	10.7
HDV	E(99.6), P(0.4)	Na/Na	Na	41.5	113	5.2
Escherichia coli	L(32.0), A(29.5), G(9.4)	LAT/GAV	50.0	58.0	765	18.7
Bacillus subtilis	A(23.8), L(19.8), S(19.8)	LKA/KSG	43.5	48.5	481	15.0
Archaeoglobus fulgidus	E(22.0), V,(18.0), L(18.0)	ER/KTL	48.6	51.9	389	10.1
Methanococcus jannaschii	E(25.9), K(22.2), L(11.1)	ILE/KGR	31.0	31.7	412	10.8
Saccharomyces cerevisiae	S(24.0), Q(18.7), N(11.7)	SQN/KDQ	38.1	44.3	759	18.5
Arabidopsis thaliana	S(27.2), G(12.3), P(11.5)	SLE/GES	36.0	50.9	812	16.0
Caenorhabditis elegans	S(14.9), T(13.8), Q(13.6)	SLQ/QGS	35.0	51.7	1103	25.0
Drosophila melanogaster	Q(31.9), A(15.2), S(11.3)	QAS/QAS	41.0	61.3	1338	15.6
Danio rerio	S(21.4), E(17.6), P(13.1)	LAG/ESK	37.6	54.4	1286	37.9
Gallus gallus	E(17.7), P(15.1), S(13.4)	LAG/ESK	50.0	62.5	1099	20.8
Mus musculus	E(19.2), P(14.6), A(11.6)	LAG/EPA	41.7	60.9	1304	26.6
Homo sapiens	E(16.0), P(16.0), A(14.3)	LAG/ESP	40.9	63.0	1390	31.2

N/C SAARs, most abundant N- and C-terminal SAARs corresponding to 5% and 95% of RCPs length, respectively; the middle point of the repeat fragments is defined as the position of repeats; GC1, genomic GC content; GC2, average GC content of repeat codon; L1, average length of RCPs; L2, average length of repeat fragments.

The positive correlation between GC content and genes rich in coding repeats has also been noticed in recent years (41–43), which suggests that the formation and evolution of coding repeats have certain constrains of the compositional specificity at the genome level. Other studies also indicate that the length of the coding sequence is directly proportional to higher GC content (44) as the stop codon has a bias toward A and T, thus the shorter the sequence the higher the AT bias (45). To investigate this, we used ProRepeat to compute the average GC content of repeats codon across taxa. The result shows that the average GC content of repeat codons is much higher than the genome GC content in nearly all species (Table 2). This is especially true in eukaryotes which have higher repeat abundance than prokaryotes and viruses. On the other hand, although the average length of RCPs is much greater than the average length of proteins in different kingdoms, i.e. 361 AA in Eukaryotes, 267 AA in Bacteria and 247 AA in Archaea, respectively (46), the relationship between GC content and the length of RCPs is not very strong.

FUTURE DIRECTIONS

To cope with the fast development of genome sequencing and annotating, we have been keeping ProRepeat updated to the latest version of the protein databases UniProtKB and RefSeq protein. Furthermore, as our repeat integrating strategy merges different datasets generated by different algorithms, we will integrate more repeat patterns into ProRepeat detected by more algorithms in the future.

Comparing specific repeat fragments among orthologous RCPs is a widely used strategy to discover their potential evolutionary and functional roles. The former analysis across *Drosophila*, rodents and primates (13,37,39,42) shows its reliability. As ProRepeat contains data over a broad taxonomy range, it may serve as an excellent platform to perform orthologous analysis on repeats. To meet such requirements, we are currently integrating ProRepeat with ProGMap—the integrated annotation resource for protein orthology (47) we developed earlier. With this setup users can compare repeats among orthologous RCPs in ProRepeat.

ACKNOWLEDGEMENTS

The authors like to thank Wilfried de Jong for starting the project together with J.A.M.L., Maarten van den Bosch for providing the prototype implementation of the database, Guido Kappé for programming the suffix tree program, Arnold Kuzniar for fruitful discussions and Jingchu Luo for critically reading the manuscript.

FUNDING

Funding for open access charge: BioRange Project SP3.2.2.2 of the Netherlands Bioinformatics Centre (financial support to H.L.).

Conflict of interest statement. None declared.

REFERENCES

1. Orr,H.T. and Zoghbi,H.Y. (2007) Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, **30**, 575–621.
2. Buchanan,G., Yang,M., Cheong,A., Harris,J.M., Irvine,R.A., Lambert,P.F., Moore,N.L., Raynor,M., Neufing,P.J., Coetzee,G.A. *et al.* (2004) Structural and functional consequences of glutamine tract variation in the androgen receptor. *Hum. Mol. Genet.*, **13**, 1677–1692.
3. Brown,L., Paraso,M., Arkell,R. and Brown,S. (2005) In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. *Hum. Mol. Genet.*, **14**, 411–420.
4. Brown,L.Y. and Brown,S.A. (2004) Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet.*, **20**, 51–58.
5. Gatchel,J.R. and Zoghbi,H.Y. (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.*, **6**, 743–755.
6. Labaj,P.P., Leparc,G.G., Bardet,A.F., Kreil,G. and Kreil,D.P. (2010) Single amino acid repeats in signal peptides. *FEBS J.*, **277**, 3147–3157.
7. Hancock,J.M. and Simon,M. (2005) Simple sequence repeats in proteins and their significance for network evolution. *Gene*, **345**, 113–118.
8. Salichs,E., Ledda,A., Mularoni,L., Alba,M.M. and de la Luna,S. (2009) Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet.*, **5**, e1000397.
9. Kobe,B. and Kajava,A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct.l Biol.*, **11**, 725–732.
10. Miller,M.K., Bang,M.L., Witt,C.C., Labeit,D., Trombitas,C., Watanabe,K., Granzier,H., McElhinny,A.S., Gregorio,C.C. and Labeit,S. (2003) The muscle ankyrin repeat proteins: CARP, ankrd2/Arpp and DARP as a family of titin filament-based stress response molecules. *J. Mol. Biol.*, **333**, 951–964.
11. Labeit,S. and Kolmerer,B. (1995) Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science*, **270**, 293–296.
12. Callis,J., Carpenter,T., Sun,C.W. and Vierstra,R.D. (1995) Structure and evolution of genes encoding polyubiquitin and ubiquitin-like proteins in Arabidopsis thaliana ecotype Columbia. *Genetics*, **139**, 921–939.
13. Simon,M. and Hancock,J.M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol.*, **10**, R59.
14. Karlin,S., Brocchieri,L., Bergman,A., Mrazek,J. and Gentles,A.J. (2002) Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl Acad. Sci. USA*, **99**, 333–338.
15. Monsellier,E. and Chiti,F. (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.*, **8**, 737–742.
16. Han,J.H., Batey,S., Nickson,A.A., Teichmann,S.A. and Clarke,J. (2007) The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell. Biol.*, **8**, 319–330.
17. Depledge,D.P. and Dalby,A.R. (2005) COPASAAR–a database for proteomic analysis of single amino acid repeats. *BMC Bioinformatics*, **6**, 196.
18. Depledge,D.P., Lower,R.P. and Smith,D.F. (2007) RepSeq–a database of amino acid repeats present in lower eukaryotic pathogens. *BMC Bioinformatics*, **8**, 122.
19. Kalita,M.K., Ramasamy,G., Duraisamy,S., Chauhan,V.S. and Gupta,D. (2006) ProtRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinformatics*, **7**, 336.
20. Katti,M.V., Sami-Subbu,R., Ranjekar,P.K. and Gupta,V.S. (2000) Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci.*, **9**, 1203–1209.
21. Faux,N.G., Huttley,G.A., Mahmood,K., Webb,G.I., de la Banda,M.G. and Whisstock,J.C. (2007) RCPdb: an evolutionary classification and codon usage database for repeat-containing proteins. *Genome Res.*, **17**, 1118–1127.
22. Robertson,A.L., Bate,M.A., Androulakis,S.G., Bottomley,S.P. and Buckle,A.M. (2011) PolyQ: a database describing the sequence and domain context of polyglutamine repeats in proteins. *Nucleic Acids Res.*, **39**, D272–D276.

23. The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
24. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
25. Gusfield,D. and Stoye,J. (2004) Linear time algorithms for finding and representing all the tandem repeats in a string. *J. Comput. Syst. Sci.*, **69**, 525–546.
26. Newman,A.M. and Cooper,J.B. (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382.
27. Alba,M.M., Laskowski,R.A. and Hancock,J.M. (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics*, **18**, 672–678.
28. Karlin,S. (1995) Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.*, **5**, 360–371.
29. Liolios,K., Chen,I.M., Mavromatis,K., Tavernarakis,N., Hugenholtz,P., Markowitz,V.M. and Kyrpides,N.C. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **38**, D346–D354.
30. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
31. Marcotte,E.M., Pellegrini,M., Yeates,T.O. and Eisenberg,D. (1999) A census of protein repeats. *J. Mol. Biol.*, **293**, 151–160.
32. Lovell,S.C. (2003) Are non-functional, unfolded proteins ('junk proteins') common in the genome? *FEBS Lett.*, **554**, 237–239.
33. Huntley,M.A. and Golding,G.B. (2002) Simple sequences are rare in the Protein Data Bank. *Proteins*, **48**, 134–140.
34. Haerty,W. and Golding,G.B. (2010) Low-complexity sequences and single amino acid repeats: not just 'junk' peptide sequences. *Genome*, **53**, 753–762.
35. Dunker,A.K., Silman,I., Uversky,V.N. and Sussman,J.L. (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.
36. Haerty,W. and Golding,G.B. (2010) Genome-wide evidence for selection acting on single amino acid repeats. *Genome Res.*, **20**, 755–760.
37. Mularoni,L., Ledda,A., Toll-Riera,M. and Alba,M.M. (2010) Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res.*, **20**, 745–754.
38. Huntley,M.A. and Golding,G.B. (2006) Selection and slippage creating serine homopolymers. *Mol. Biol. Evol.*, **23**, 2017–2025.
39. Huntley,M.A. and Clark,A.G. (2007) Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. *Mol. Biol. Evol.*, **24**, 2598–2609.
40. Zhang,L., Yu,S., Cao,Y., Wang,J., Zuo,K., Qin,J. and Tang,K. (2006) Distributional gradient of amino acid repeats in plant proteins. *Genome*, **49**, 900–905.
41. Nakachi,Y., Hayakawa,T., Oota,H., Sumiyama,K., Wang,L. and Ueda,S. (1997) Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. *Mol. Biol. Evol.*, **14**, 1042–1049.
42. Alba,M.M. and Guigo,R. (2004) Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.*, **14**, 549–554.
43. Cocquet,J., De Baere,E., Caburet,S. and Veitia,R.A. (2003) Compositional Biases and Polyalanine Runs in Humans. *Genetics*, **165**, 1613–1617.
44. Pozzoli,U., Menozzi,G., Fumagalli,M., Cereda,M., Comi,G.P., Cagliani,R., Bresolin,N. and Sironi,M. (2008) Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol. Biol.*, **8**, 99.
45. Wuitschick,J.D. and Karrer,K.M. (1999) Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila. *J. Eukaryot. Microbiol.*, **46**, 239–247.
46. Brocchieri,L. and Karlin,S. (2005) Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.*, **33**, 3390–3400.
47. Kuzniar,A., Lin,K., He,Y., Nijveen,H., Pongor,S. and Leunissen,J.A. (2009) ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Res.*, **37**, W428–W434.